DOCUMENT RESUME

ED 263 122 TM 850 556

AUTHOR McLarty, Joyce; And Others
TITLE Using Multiple Data Sources in Teacher Evaluation.
PUB DATE Apr 85
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL. April 1-4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Observation Techniques; *Data Collection; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Evaluators; Interviews; *Occupational Tests; Peer Evaluation; Portfolios (Background Materials); Questionnaires; State Programs; Student Evaluation of Teacher Performance; Teacher Behavior; *Teacher Evaluation; *Teacher Promotion; Teaching Skills; Test Construction
IDENTIFIERS *Tennessee Career Ladder Program

ABSTRACT

The Tennessee Career Ladder Evaluation system, which uses a multiple data source approach to assessment, was developed to identify excellent teachers. It is used to determine whether or not the teacher should receive increasing career benefits. Evaluation data are generated by three evaluators, the teacher candidate, the school principal, three peer teachers, and twenty or more students. Extensive review was used in generating six categories of teacher competency: planning; instructional strategies; evaluation, classroom management; leadership; and effective communication. Items were field tested in different types of data collection: classroom observation; interviews with teachers and principals; questionnaires for students and peer teachers; portfolio rating; a written test; and evaluator judgement. Results from each instrument were analyzed separately and then reanalyzed as part of the set. Intercorrelations were low, confirming that each source measured different viewpoints or aspects of performance. The resulting scores were reduced to a smaller number of scores, and then weighted. It was concluded that the multiple data source system was difficult and expensive to build, relatively inflexible, and complicated to explain. However, it provided a thorough and equitable evaluation, was relatively stable, and was logical. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

ED263122

USING MULTIPLE DATA SOURCES IN TEACHER EVALUATION

Joyce McLarty

Carol Furtwengler

George Malo

State Department of Education

State of Tennessee

presented to

the National Council on Measurement in Education

Chicago, 1985

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

McLarty, Joyce

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 556

Tennessee's Comprehensive Education Reform Act of 1984 provides for the development of a Career Ladder system through which meritorious teachers may receive added recognition and pay. The Career Ladder Evaluation System is the summative evaluation system used to determine whether or not the teacher should receive Career Ladder benefits above the basic (Career Ladder I) level. Currently employed teachers with 8 years of experience are eligible to apply for Career Ladder II, and those with 12 years of experience are eligible to apply for Career Ladder III. Whether or not they qualify for those levels depends entirely on the results of the evaluation.

The Tennessee Career Ladder evaluation system uses a multiple data source approach to assessment. This paper describes the multiple data source approach as it was developed and field tested in the Tennessee program, the problems of developing and implementing the multiple data source approach, and the advantages and disadvantages of such a system.

Multiple Data Sources

The multiple data source approach to assessment utilizes several sources of information in the assessment process. These sources may be different respondents, different data collection formats, or both. In the Tennessee program, both are used. Respondents include the evaluators (three of

them), the teacher candidate, the school principal, peer teachers (three of them), and students (at least 20 of them). Data collection formats include the interview, portfolio rating, written test, classroom observation, and questionnaires.

The Tennessee Career Ladder Evaluation System is a competency based assessment. In a multiple data source system, this can be handled either by assigning a specific competency to a single data source (eg. classroom management to the observation and planning to the portfolio) or by assigning competencies to more than one data source. In the Tennessee system, each competency is assigned to at least two data sources. This provides for an assessment of the desired teacher performance which is less affected by the particularities of a specific respondent or data collection method. It protects the candidate against the biases of a particular principal. It compensates for the teacher who may have great difficulty in conducting a class when the evaluator is present for reasons unrelated to his or her performance without the evaluator in the room. The multiple data source system provides counterbalances as protection against variance associated with respondent and data collection methods.

The data sources are distributed as follows:

Type: Respn:	Observation	Interview	Questionnaire	Portfolio	Test	Judgment
Eval	X	X		X		X
Cand		X		X	X	
Princ		X				
Peer			X			
Stud			X			

Competencies

In order to select competencies to be assessed, hundreds of statements related to effective teaching or teacher practices were generated through examination of the research literature. These were circulated to Tennessee teachers for review and comment. After revisions based on teacher input, the competencies were again circulated, this time for approval. Those competencies accepted by most of the reviewing teachers were retained. From these, a subset of measurable, generic, and important competencies were identified. These were grouped and ordered hierarchically with changes in wording as needed. The final competencies which were approved are organized into five domains: planning, instructional strategies, evaluation, classroom management, and leadership. These are subdivided into indicators and then into measurement items. In addition to

the five domains and their indicators, a sixth domain, Communicates Effectively, was identified containing indicators which would be required at a minimum competency level but would not contribute to the Career Ladder status. The four indicators were performance of professional responsibilities, reading, writing, and oral communication.

Development of Data Sources

Items measuring as many of the indicators or measurement items as possible were developed for each data source. Development procedures varied according to the type of instrument but always included item review and pilot testing followed by item revision and additional pilot testing as needed. As field tested, the Tennessee Career Ladder Evaluation System used a structured interview of the candidate (conducted by one of the evaluators), a brief structured interview of the school principal (conducted by a second evaluator), a rating of the candidate's portfolio (by the third evaluator), three classroom observations (one by each evaluator), completion of questionnaires by one class of the candidate's students, completion of questionnaires by three of the candidate's peers, and a written test taken by the candidate. After the field test, the principal interview was converted into a questionnaire.

Classroom Observation

Coding of teacher behaviors related to the competencies formed the basis of the observation system. All coding of the observation form began with the teacher's action. Although the teacher action column was divided into sections for ease of coding, the specified activities did not need to occur in order. These sections were task description, presentation of content, practice and review, monitoring and adjustment, providing feedback, and other classroom procedures. This approach permitted the evaluator to begin coding at any stage in the lesson, and also allowed full credit to teachers using instructional methodologies (e.g. direct instruction, discovery methods) which require different sequencing of activities.

In addition to space for coding teacher activities, the observation form included space to code teacher affect (verbal or non-verbal), and student response or student initiated activities (appropriate or inappropriate). After pilot testing, this area was reduced to include only teacher behavior, teacher positive affect, student response, student initiation, and the appropriateness of student response.

A second part of the observation form was developed to handle classroom management data. Coding of this began when a student or students were noted to be off-task. Teacher response (including no reaction) and the resultant student behavior (compliance or continued off-task behavior) were coded. Numbers of student academic and procedural questions

were also tallied. Finally a series of direct questions was included. This series included the general teaching method in use (lecture, discussion, independent work), the group size (whole class, subgroup, individuals), method of selecting students to respond to questions, use of media or materials, extent of student lack of understanding, teacher reteaching in response to lack of understanding, numbers of students (in the class, off task, waiting), interruptions occurring during the observation period (avoidability, elapsed length between end of interruption and resumption of class), and evidence of favoritism on the basis of race or sex (in academic interactions, discipline, or positive affect).

The two coding areas of the observation form were combined to form a "scan" of four minutes. Three additional minutes were allotted to the evaluator to complete the added questions. Seven of these scans were to be completed in a typical observation period. This set was preceded by a pre-observation conference with the candidate in which the evaluator was to be briefed on the intended content of the class (nature of the lesson, subject and grade level, major objectives, activities, and procedures for evaluating student learning) and any special circumstances such as students of low socio-economic status, one or more mainstreamed special education students, slow learners, gifted students, lack of instructional materials, teacher assignment to teach an inappropriate course or grade level,

one or more students with special problems, and organizational dysfunctions).

After the observation, the evaluator answered additional questions about the teacher's communication skills (speech, grammar, writing), planning and management, use of resources, and facilities. A post-observation conference permitted the teacher to explain any aspects of the class which might otherwise be incorrectly interpreted by the evaluator. In it, the evaluator also shared one strength and one weakness noted with the teacher.

The observation system used a low inference approach. That is, the evaluator was asked to record behavior rather than an evaluation of that behavior. In extensive training, the evaluator was taught definitions and categories and which specific behaviors were to be marked where on the form. If the specified behavior occurred, it was marked. There was no tallying of the number of times a specified behavior occurred. Instead, the presence or absence of the behavior was recorded in each of a series of four minute scans. These were later aggregated over the observation sets.

While this observation format is not unique among low inference observation systems, the method of selecting the behaviors to be observed appears to have been. The form and, therefore, the evaluator's attention were focused on behaviors which related to the specified objectives to be measured. There was no attempt to observe, like a camera,

everything which occurred in the specified time and location. Activities which were unrelated to the objectives were simply ignored. If there was no obvious way of scoring a behavior as evidence for one of the objectives, it was not recorded. The same was true of fine distinctions in the behavior of either the students or the teacher. The only two reasons for differentiated recording of behavior which were accepted in the development process were (1) the differentiation made the observation easier or more reliable, or (2) the differentiated behaviors needed to be scored differently.

Candidate Interview

The candidate interview was developed as a highly structured series of basic and follow-up questions to be asked of each candidate. Each of the question sets was accompanied by a behaviorally-anchored five-point rating scale. The rating scales were cumulative with the lowest level indicating competence below the Career Ladder I level, and the highest indicating exemplary performance on it. Candidates were asked to provide documentation to accompany their answers to the interview questions to make the interview process more valid and reliable. The initial version of the interview took more than three hours to conduct. Rather than reduce the number of questions per rating and, therefore, the thoroughness of the rating, entire question sets were deleted from the interview process prior to the field testing. Those kept were selected on the

basis that they were not as well covered by alternate data sources.

Superordinate (Principal) Interview

More like a questionnaire than an interview, the principal interview was to be administered directly by the evaluator. This process was to minimize the time requirement for the principal (who might have a dozen or more teachers applying for the Career Ladder) while maintaining direct contact between the principal and evaluator and allowing the principal to add orally information which he or she might feel unwilling to commit to paper. Ratings were on a four-point scale ranging from below average to outstanding. The principal could also indicate that he or she did not know the answer and designate someone else to respond to that question.

The questionnaires did not ask for opinions or attitudes, but rather sought to extend the observation system. This approach was designed to minimize the influence of respondent preferences, and focus the respondent's attention on observable attributes of teacher behavior. The principal interview did not ask for the principal's opinion of the teacher, but rather used the principal to extend the observation system in the areas delineated by the competencies. Each item asked for information about the teacher's behavior (e.g. develops student remediation plans which are based on diagnostic data).

Peer-Teacher Questionnaire

The peer-teacher questionnaire was designed to extend the observation rather than to solicit opinions. Only competencies in the leadership domain were considered suitable for inclusion in the peer-questionnaire since peers do not generally observe each other in the classroom. Typical content of the peer questionnaire asked about whether this teacher generally shared materials and ideas with other teachers.

Student Questionnaire

Like the principal interview and peer-teacher questionnaire, the student questionnaires extended the observation system. Two forms of the student questionnaire (one for grades 2-6 and the other for grades 7-12) were developed. Differences between the two forms were primarily in the level of vocabulary used and the length. The secondary form was developed first, then shortened and simplified for elementary use. Typical content for the student questionnaire asked whether this teacher generally noticed when students were not paying attention.

Written Test

Two of the screening competencies (that those which were evaluated as pass or fail rather than as contributors to the Career Ladder level determination) were ability to read and write. These were best measured by minimum competency written tests. A multiple choice reading test, suitable for

teachers, administrators and other certificated educators was developed according to criterion-referenced measurement procedures and a cut-off score set for it. A set of writing sample prompts and scoring procedures for holistic, pass or fail scoring, were also developed for field testing.

Since four of the domains (planning, teaching strategies, evaluation, and classroom management) had knoweldge components, a criterion-referenced multiple-choice test was developed to cover these. Data from this Professional Skills test was handled in the same way as data from the other instruments. No cutoff scores were used with it for Career Ladder II or III status determination.

Portfolio Rating

The Tennessee Career Ladder portfolio, like the other instruments was keyed to the competencies. Each item identified for inclusion was related to one or more of the domains. Teachers were asked to submit a set of instructional plans (unit and lesson); indicators of student progress; classroom rules, policies, and procedures; a list of professional development activities; and a list of leadership activities. Each portfolio rating area was accompanied by a five-point rating scale. Some of these scales were cumulative in nature, like those in the interviews. Others were collective (scales in which a series of indicators were evaluated separately then summed for a rating. For example, if four of the five indicators were noted, the rating was a four). Instructions to the

candidate and evaluator specifically discounted the appearance of materials (the "stained glass" effect) in favor of the quality of the content. Candidates were asked to focus on showing competency and not merely compiling masses of data.

Evaluator Judgment

In addition to administering the formal instruments, all of which were highly structured, the three evaluators contributed to the final determination of the candidate's Career Ladder status with a consensus judgment. This constituted an eighth data source.

Field Testing

Once each of the instruments had been separately developed, they were field-tested as a set using the same field test candidates and the same evaluators. Field test candidates were nominated as a stratified random sample of K-12 teachers in the state. The stratification was based on grade level, subject area taught (for junior high and high school), and location. Each of the 142 systems in the state had at least one teacher nominated as a field test candidate, with the remainder allocated proportionally by the size of the system. Field test candidates were paid $100 each for their participation. This was because field test candidates had to commit several days to the process: they were observed three times, interviewed at length, and tested for five hours. Their principal was interviewed, and

their students and peer-teachers queried. They had to prepare interview and portfolio materials. The evaluation, in some respects, was more like a case study.

Evaluators were classroom teachers nominated as outstanding. Individually selected, they had to be able to learn a new and complex evaluation system, deal appropriately with teachers and administrators, and travel statewide alone for a period of more than a month. Evaluators had to be able to administer each of the instruments reliably. It took more than two weeks, 8-10 hours a day, to train the evaluators for the field test (the training of the evaluators has since been lengthened to four weeks). While no one had difficulty learning to administer questionnaires, the interview, portfolio, and observation required considerable (and to some extent different) skills. Each evaluator had to master them all.

Although the multiple-data source approach was more difficult than, for example, separate field tests for the different instruments using different field test samples and different evaluators, it had some major advantages as well. Candidates subjected to all of the instruments had, indeed, received a very thorough, highly objective evaluation. They had been visited on three different occasions by highly trained evaluators who were outstanding classroom teachers of the same or similar subject areas or grade levels. Not only had they been observed teaching, but they had also been given a chance to talk about it, to show materials they had

prepared, and to receive feedback from three different evaluators. Their principal, three peer teachers, and their students had been consulted.

The evaluators, having begun the field test feeling concerned about their ability to be fair and impartial to candidates, returned from the field test with few remaining concerns. They reported that they felt comfortable with the objectivity and validity of the instruments, and secure in their ability to administer them properly. Some reported that they had learned more during the field test, through their focused attention on specific teacher and teaching attributes, than they had in their own teaching careers. They considered the extensive training and experience as an evaluator outstanding staff development.

Data Analysis

As with the rest of the field test, data analysis was complicated by the multiple data source approach. Each instrument was analyzed first separately (all were highly reliable), then reanalyzed as part of the set. Low intercorrelations indicated that the different sources of data were, as intended, picking up different viewpoints on or aspects of performance. With over 300 separate scores for each candidate, the evaluation gave a very detailed profile of each teacher's performance. In fact, the detail was so great that it had to be reduced in order to be understandable. Scores were, therefore, combined within the same instrument and competency to generate more meaningful

information. Because of the variety of scales used in the individual instruments (zero to one, one to one hundred, one to four, one to five, etc.), the scores also were rescaled to the same scale. A nine-point stanine-like scale was chosen, which also served to normalize somewhat the severely skewed data for some of the scores. Eighty-two scores were converted to this scale.

Placing the 82 scores on the same scale made it possible to combine them to get total scores, not by instrument but by competency. The step was logically sound. If a candidate did well or poorly, it should have been on the competency rather than on the instrument. The structural integrity of the multiple-data source system allowed the focus to be on the competency rather than on the method by which it was measured.

In order to combine scores by competency, it was necessary to weight them. Each of the 82 scores received a weight based on the importance of the competency it measured, the number of items used in creating it, and the level of knowledge and objectivity of the respondent supplying the information. Each domain was weighted and combined separately to a total of 100% (100 points). This data-based score was in turn weighted at 80% and added to the evaluator consensus rating weighted at 20% (again, for each domain separately). No totals across domains were used.

Instead, cutoff scores were set for Career Ladder II and III for each of the five domains separately. Candidates needed to reach the designated level for Career Ladder II in each of the five domains to qualify. For Career Ladder III, the candidate needed to reach the designated level in four of the five domains and the level of Career Ladder II in the remaining area.

Implementation

The difficulty of the technical problems in combining scores across instruments in order to measure competencies was exceeded only by the difficulty of explaining how it was done. The processes of aggregating scores, rescaling them, weighting and combining them, and then reweighting and recombining them to allow for evaluator judgment was confusing to almost everyone. Once the system was implemented and the evaluation results had status and monetary implications, explaining the system to the candidates became a major concern. The kinds of questions candidates really wanted answered represented concerns such as how much the observation was worth. Candidates were unsatisfied by the answer that the weight depends on the domain. The nine-point scale added to the confusion. It was reported as being an attempt to inject a quota into the evaluation system. Explanations that the scale applied only to normalizing the eighty-two individual scores, and that cutoff scores were set prior to any actual evaluations,

thereby, preventing Career Ladder decisions based on quotas were not accepted.

What candidates appeared to want most to know was exactly what they had to do - what behavior, how many times - in order to qualify. They wanted to know what "percent correct" was needed to "pass," and the complexity of the system made that difficult for them to see.

It also made the system difficult to manipulate. It becomes hard to stage the highest-scoring possible classroom lesson if you can not determine whether it is more valuable to have students carry on an animated discussion or to give them an exemplary lecture. The way the multiple data source system is balanced, excessive behavior in any direction (eg. all discussion and no presentation, or all presentation and no discussion) is likely to receive a lower score. The exact proportion of discussion to presentation which would earn the highest score, however, is neither specified in the manual nor easily calculated.

That is frustrating to those who would willingly change their teaching behavior in order to qualify for the merit increments. They argue that if the Career Ladder system is to improve instruction, such specificity is needed in order that teachers modify their teaching practices appropriately. The primary purpose of the Career Ladder evaluation system is the identification of outstanding teachers. Providing information at a level of specificity which would permit less than outstanding teachers to appear to behave in the

same way as outstanding ones, at least for short periods, would diminish the validity of the summative evaluation system.

If the multiple data source evaluation system is difficult to manipulate for inappropriate purposes, it is also difficult to manipulate for appropriate ones. A candidate who receives a low score in only one competency area may make significant and effective improvements in that area. In order to reevaluate the candidate in that area, however, several (if not all) of the instruments have to be readministered since they all contribute to the competency-based score. For the same reason, incomplete evaluations (those missing an instrument) cannot be processed. If the principal is unable or unwilling to complete the questionnaire, what can be done to replace those scores? What if the students cannot respond to a questionnaire (e.g. special education students)?

Carrying an evaluation over from one school year to the next is also problematical. Is an evaluation begun while the teacher worked with one classroom of students appropriately completed with another? Suppose the teacher (or the principal) changes schools. Who fills out the principal questionnaire, the old principal or the new one? With dollar amounts of $2000 and up per year depending on the results of each evaluation, these questions are far from academic.

On the positive side, the multiple data source system is relatively resilient. When the Tennessee Education Association released condidential rating scales for the interview and portfolio to selected teachers, revision to the system to mitigate the effect was accomplished in a few days. The interview was dropped altogether, and weights were reduced for affected portfolio items. Each domain was adjusted by assigning the weights from the interview and affected portfolio items to the other instruments measuring the same competency.

Conclusion

It is too soon to identify all of the advantages and disadvantages of Tennessee's multiple data source evaluation system. It has been implemented only since fall, 1984, and the first Career Ladder II and III teachers have yet to be identified. It seems safe to say, however, that the multiple data source system is difficult and expensive to build, relatively inflexible, and complicated to explain. It also provides a thorough and equitable evaluation, is relatively stable, and is based on concepts which are logically appealing.